On the Optical Character Recognition and Machine Translation Technology in Arabic: Problems and Solutions

Prof. Oleg Redkin, Dr. Olga Bernikova
Faculty of Asian and African Studies, St. Petersburg State University, St. Petersburg, Russia

Abstract - *The report addresses the basic problems of the Arabic language formalization based on analysis of linguistic errors in software products. Reviewing the principles of modern information systems operation the authors come to the conclusion that the existing methods of the Arabic formalization allow to note a shift towards the technological aspects of the linguistic processing of facts, however, the quality of applied linguistic components still remains poor. Possibilities for the application of traditional recognition algorithms for Arabic are still uncertain in spite of a significant number of theoretical and practical results in the field of computational linguistics. There are several problems which are due to be solved in relation to the processing of the Arabic text. These issues may be divided into those related to Optical Recognition of the Written Text (OCR), Word Processing (WP) and building of the content of the dictionaries, Machine Translation (MT).*

Keywords: NLP, Arabic, OCR, machine translation.

1 Introduction

Modern tendencies of developing in the information-oriented society can be defined by the deep interaction of the international industrial, technical, economic, sociological, scientific and informational aspects. Globalization in every aspect of everyday life is primarily performed through language, which is the main tool of communication. Expansion of cross-cultural interaction on different levels calls for the necessity to eliminate the so-called "language barrier"; this can be achieved through the application of the machine translation systems and improvement of the multilingual search engine software.

Recently, the progress in development of quantity and quality of the software applications for processing of linguistic material in Arabic became evident. However the available linguistic software products still abound in errors and obvious mistakes. This largely happens due to the intention to shift the focus towards the creation of the technological solutions for the universal processing of linguistic material, while the quality of the linguistic background tends to be left in a poor condition.

2 Historical background

The commencement of the language formalization process, aimed at establishing of the automatic parsing of linguistic material dates back to early 60's[1]. However just in 1981 John McCarthy generalized the theory to deal with the conjugational system of Arabic, on the basis of an autosegmental account of vowel and consonant slots on a central timing tier (A CV skeleton analysis of Arabic root-and-pattern morphology) [2].

The first two-level morphological analysis was carried out in 1983 by K. Koskenniemi. At that time he presented a new linguistic, computationally implemented model for morphological analysis and synthesis. It was general in the sense that the same language independent algorithm and the same computer program can operate on a wide range of languages, including highly inflected ones such as Finnish, Russian or Sanskrit. The model was based on a lexicon that defines the word roots, inflectional morphemes and certain nonphonological alternation patterns, and on a set of parallel rules that define phonologically oriented phenomena. The rules are implemented as parallel finite state automata, and the same description can be run both in the producing and in the analyzing direction[3]. His system was based on the prefix-suffix method of word formation and a simple sequence of morphemes. The changes of the internal structure of the root, as well as the conversion of weak roots were not considered at that time.

In 1996, Xerox Research Center has implemented a system using a combination of automated algorithms for root structures and derivational patterns. This technology is proved to be very effective for the linguistic data processing, and now it is used in NooJ - a linguistic environment, processing the texts of a few million units in real time. It includes tools for compiling, testing and supports the bulk of lexical resources, as well as morphological and syntactic grammars[4]. However, our analysis of the NooJ operation shows that the proposed description of the principles of the Arabic language formalization is not comprehensive and, as a result, does not provide the sufficient accuracy of data processing. In

[1]Bar-Hillel, Y., "The Present Status of Automatic Translations of Languages"; Advances in Computers, Vol.1, 91-163, 2006.
[2]McCarthy John J. "A prosodic theory of nonconcatenative morphology"; Linguistic Inquiry, Vol. 12, 373-418, 1981.
[3]Koskenniemi, Kimmo, "Two-level Morphology: A General Computational Model for Word-Form Recognition and Production, Publications", University of Helsinki, Department of General Linguistics, 1983. P.160.
[4] http://www.nooj4nlp.net/pages/nooj.html.

particular, insufficient quantity of verbal derivational algorithms was noted along with the lack of the complex solutions for the root definition based on various patterns of broken plurals, distinctions in the final hamza writing when joining enclitic, etc. Thus, the paper "Standard Arabic Formalization and Linguistic Platform for its Analysis" by Slim Mesfar, describes NooJ as «system that uses finite state technology to parse vowelled texts, as well as partially and not vowelled ones. It is based on large-coverage morphological grammars covering all grammatical rules»[5]. The author highlights that for Arabic, as a highly inflectional and derivational language, they had to define three new operators such as the <T> operator that checks if the last consonant within a noun is a " ة" (T - Teh marbouta) to replace it with a " ت" (t –Teh maftouha) in some inflexional or derivational descriptions:

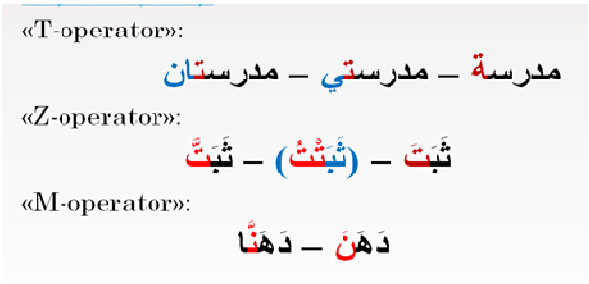

Figure 1. Special "operators" for Arabic in accordance with "Standard Arabic Formalization and Linguistic Platform for its Analysis" by Slim Mesfar.

Noting that the system contains all grammatical rules the author misses a number of grammatical features of the Arabic language. For example, there are no "operators" for broken plurals as there is no distinction in types of plural patterns, depending on the proper part of speech:

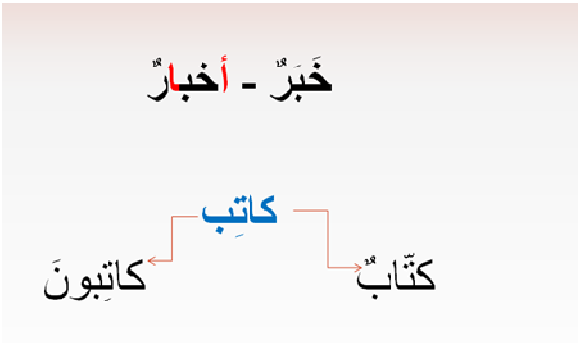

Figure 2. Distinction in plural patterns, depending on the part of speech and the example of broken plural pattern.

Besides that, Slim Mesfar notes that "by inflectional description it was composed the set of all possible transformations which allow us to obtain, from a lexical entry, all inflected forms. These inflexional descriptions include the mood (indicative, subjunctive, jussive or imperative), the voice (active or passive), the gender (masculine or feminine), the number (singular, plural or dual) and the person (first, second or third). On average, there are 122 inflected forms per lexical entry"[6]. According to our research there can be up to 400 forms and, taking into consideration the possibility of enclitics joining for the transitive verbs, up to 780 models.

Our findings in this area are based on experience in the development of the morphological analyzer which has the following characteristics:

- More than 60000 rules for verbs.
- More than 7500 for names.
- The rules do cover all kinds of roots.
- Lexical content includes 20000 units.
- All distinctions in *hamza* and *tashdeed* writing are taking into consideration.

3 Problems

The recent developments in the field of the Information and Communication Technologies open new horizons and perspectives for scholars engaged in the field of linguistic analysis and linguistic data processing. It covers a wide range of areas and allows to create new programs based on the Arabic script such as data bases, automatic translation and automatic Arabic text recognition, search engines in the Internet, teaching tools, and, finally, Arabic-Arabic and Arabic-multilingual dictionaries.

Along with the advantages brought by the common development in the field of Arabic software there is a number of difficulties. One of the most important of them is the proceeding of the Arabic text. In order to cope with these challenges we have to develop new principles and ways, which aim to solve the problem of the Arabic language text formalization.

There are several problems which are due to be solved in relation to the processing of the Arabic text. These issues may be divided into those related to Optical Recognition of the Written Text (OCR), Word Processing (WP) and building of the content of the dictionaries, Machine Translation (MT).
Unlike Latin or Cyrillic scripts, the Arabic one brings along many peculiarities, which cause many difficulties the researches confront with.

Such difficulties are represented by a wide variety of Arabic fonts (the major are kuufii, deewaanii, ruq'a, naskh, thulth, etc.) as well as by elements of individual letters, the various forms of writing of several letters depending on their

[5] Mesfar S. "Standard Arabic formalization and linguistic platform for its analysis"; Proceedings of the Conference The Challenge of Arabic for NLP/MT, London, England, 2006. P.84.

[6] Mesfar S. "Standard Arabic formalization and linguistic platform for its analysis"; Proceedings of the Conference The Challenge of Arabic for NLP/MT, London, England, 2006. P. 86.

position in the word, and ligatures, symbols for germination and vowel signs (diacritics), which are spelled sporadically.
Besides their graphic representation, Arabic letters are distinguished by dots. These dots may be placed above or below letter characters, in hand-written texts and manuscripts the precise location of these dots may vary.
Besides that, there are special characters used in local dialects (i.e. ڤ , پ , ڽ , ڜ , چ), or characters designated for sounds which are not typical to Arabic (ڤ , چ , پ , etc.). should not only be taken into consideration but considered as vital elements.

Along with the vast variety of fonts and handwritings, as well as the specifity of realization of the Arabic letters due to their position in the word, there is a big number of possible written variants of almost every Arabic word. These variants depend not only on the grammatical categories of definite / indefinite article, case (nominative, genitive, accusative), number, but on a written realization of some graphic elements, for example, dots under "yā' - ي ", hamza ء, madda آ, waşla ٱ, tashdeed ّ, etc., i.e. the elements which have a 'facultative' character in modern Arabic texts, especially in printed ones.

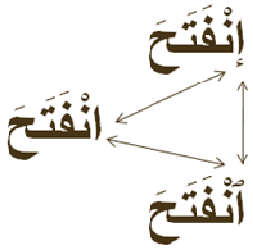

Figure 3. Examples of differences in hamza ء writing.

Even the very preliminary analysis of the Arabic text reveals the difficulties of the automatic morphological classification. The difficulties in formalizing of the Arabic language are due to the vast variety of the possible written representations of words and word forms. So, the number of the word variations derived from one stem is very significant.

Thus for maşdars, participles, and adjectives and nomina loci et temporis, and other nouns the great variety of forms is resulted by various prefixes and suffixes, inseparable prepositions, inseparable conjunctions, terminations of declension and broken plurals, etc. The variety of forms is also caused by the infixes of the eighth stem and different morphological models of the broken plurals.

Adding to this paradigm the possible variants of the written forms (with or without hamza ء, tashdeed ـّـ , waşla ٱ and madda آ), and not taking into consideration the vocalization structures, which actually reduces the total number of the forms, the list of written variants is very vast. For example, there are around 200 written variants of the word باب 'door'.

As for the paradigm of the verbal forms, in comparison with the noun ones the total number is much higher due to the more complicated and developed system of the verb conjugation in Arabic. Adding to this three grammatical categories of numbers – singular, dual and plural, active and passive voices, two genders, verbal forms in the first, second, and third persons (in dual – the second and the third persons), and, finally, five moods, one can imagine the huge spectrum of the verbal paradigm in Arabic.

Thus, the number of the forms only in the active voice of the first form (perfect and imperfect, five moods) is as high as 76. As for the graphical realization the optional representations are possible (with or without hamza, tashdeed, waşla and madda). Besides that, the inseparable particles increase the number of possible written forms of the subjunctive, jussive, energetic and imperative moods due to the number of the particles. The vocalization increases the total number of the derivates and may double it.

The total number of the verbal derivates in all possible stems may include several thousands and, with optional variants, the total number is up to 10000 forms. For example, the verb مرّ 'to move' has more than 12000 possible written derivates, including all possible stems and optional written variants (without vocalization)

Thus, the search engines based on the alphabetic principle such as in Indo-European languages, in Arabic in many cases prove to be ineffective.

4 Optical Character Recognition (OCR)

All the mentioned above incredibly multiplies the number of the written variants in Arabic and, as a consequence, difficulties of the Arabic OCR. The lack of precise standards of written Arabic along with the multiplicity of the grammar forms lead to complex problems in classification of the images of separate characters and words. It is necessary to develop new attitudes and solutions which would be valid within the framework of multivariate and multicriteria analysis, and which will be adapted for the peculiarities of a certain text.

Traditional algorithms of OCR for Arabic still remain insufficiently developed in spite of a big number of theoretical and practical achievements in the field of computational linguistics. The character recognition problem of Arabic text is more complex than that one in terms of texts based on Roman or Cyrillic alphabets.

Effectiveness of existing character recognition systems for Arabic to a great extent depends on the structure of a specific text and may vary within a very significant range.

In many cases the search engines based on the alphabetic principle, which proved to be effective enough for the Indo-European languages, are ineffective for the Arabic language.
Meanwhile, the systematic description and automatic analysis of the data base of the Arabic texts would provide the creation of the empiric principles of character recognition and search engines in Arabic. Researches of this kind have been

successfully carried out for several years at the Arabic Chair of St. Petersburg State University, Russia.

The results of OCR besides the basic characteristics of computer soft- and hardware depend on the quality of the text, the font size and its type, quality of paper, type of printing, printer ink, colors etc.

Figure 4. The sample of the vertical text segmentation in accordance with "Automatic Recognition of the Printed Arabic Texts", Shalymov D.S.[7]

The OCR of handwritten texts, in which the variation of Arabic letters is higher, is even more difficult. For instance, there is possibility of the displacement of the diacritics (dots) with respect to key elements of alphabetic characters in handwritten texts. Since the Arabic alphabet has 18 letters which differ by the number of dots and their positions (above or below the character), the dots may be considered as one of the most relevant elements in the OCR process. The question is twofold: 1) how to define the possible spectrum of the variety of the key elements of certain characters and how to define the borders of the of the position options of the elements against each other? In other words: what is the permissible scale of the transformation of characters in Arabic and dispersion of the elements which compose the characters? And how far from the main character element the dots can be located and what are the possible shapes of the dots that allow the optical recognition of the word?

Since the perception and the recognition of a certain character and combination of characters is a cognitive process based on analysis the graphic environment within a word, and, finally, the use of the graphic, morphologic and semantic contexts of the sentence or the entire text.

The cognitive process of the analysis can be phased (recognition of a character, recognition of a character within the context of a word, recognition of a character within the context of a word -(sentence) text), or take place simultaneously.

In computer analysis each stage requires real-time data processing. The amount of the data depends on the level of analysis – i.e. analysis of the optical data on the initial stage, optical and morphologic and semantic data on the word level, optical-morphologic-semantic syntax data within the context of a sentence, etc.

The huge amount of data to be analyzed demands new attitudes and approaches in building mathematical model and simulators of this process.

[7] Shalymov D.S. "Automatic Recognition of the Printed Arabic Texts"; Stochastic Optimization in Computer Science, Vol.3, 2009. P.125.

The solution may be found on the basis of the technology of simultaneous perturbation stochastic approximation algorithm (SPSA) developed by Prof. O.Granichin [6], [7] and Dr. D.Shalymov [18] - the methods which can be applied not only to Arabic but to for other languages as well.

4 Machine Translation

The thorough user-oriented study of the automatic translation systems available today and operating online with a free access has demonstrated the significant growth in the number of translation resources of Arabic. Initial testing of these translation platforms allowed us to determine several systems that carry out the most accurate translation. As the primary resource platform we will select http://translate.google.com. This choice has been created due to several factors. First, we wanted to analyze the curve of the system operation improvement over the past 5-6 years. Since the materials regarding errors in the system (http://translate.google.com operation) from 2006 became available to us, it was interesting to review the translation of the same model in a diachronic and analyze the process of the linguistic components improvement. On the other hand the resource like http://translate.google.com is one of most popular tools for on-line automatic translation, addressed by thousands of Internet users daily.

The following resources were selected as the objects for selective comparison of incorrect translations: http://www.apptek.com, http://www.translation.babylon.com, http://www.systranet.com.

This study aims to identify the most typical linguistic misinterpretations and errors that occur in the leading systems of automatic translation from Arabic into English and Russian. The role of English in this context is considered through the prism of its frequent use as an intermediate language ("hub-language").

The analysis of the quality of automatic translation from Arabic into English and Russian leads to the following conclusions, based on typical errors:

1. Earlier we have already noted that the allowable variation in writing hamza creates difficulties in the formalization of the Arabic language which is certainly reflected in the work of computer-aided translation.

Furthermore, a series of examples demonstrates the fact that the system cannot find the grapheme و (waw), acting as an "and" if the next word starts with أ:

وأكد ياكونين الذي تحدث عن نتائج زيارته الأخيرة الى الولايات المتحدة أن شركته مهتمة بالتجربة الأمريكية في هذا المجال

«The Yakunin, who spoke about his recent visit to the United States that his company is interested in the American experience in this area».

«Якунин, который рассказал о своей недавней поездке в Соединенные Штаты, что его компания заинтересована в американский опыт в этой области».

This misinterpretation takes place due to an error of the system (Google translation) algorithms which are only partially represented by a combination of words containing the initial Hamza. It is interesting to note that when testing the other above mentioned translation systems the given problem is mostly preserved.

Thus, the modern systems of automatic translation should enhance the quality of its operations by means of the creation of the rules for all possible variants of the hamza writing.

2. Another complex problem of the machine translation systems corresponds to the morphological system of Arabic.

The two important aspects are necessary to be identified here.

Firstly, the correctness of the morphological values transfer in the Arabic language is largely dependent on the specifics of its writing system. Thus the absence of vocalizations in the letter leads to the disappearing of some markers of inflectional patterns, which in its turn leads to emerging of errors in the translation. Secondly,, the morphological paradigm of the Arabic language is only partially presented in the modern computer-aided translation.

Thus the example below demonstrates that the system does not translate the verb in the feminine, subjunctive mood. While the translation of the same word in the masculine is given.

فقال الزوج : لو لم تفتح فمها ما وصلت إلى هذا
المصير، لذلك من الأفضل لك أن *تلتزمي* الصمت

«Муж сказал: не открыть рот, если вы к этому судьба, так что это лучшее, что вы *Tltzme* молчания».

The other example is the incorrect translation of the future tense (the suffix wasn't considered):

وأضاف أن الطرفين *سيوقعان* قريبا على مذكرة تفاهم بهذا الشأن تضم أيضا التعاون في إنشاء مراكز لإجراء اختبار الابتكارات الجديدة في مجال صنع القطارات

«Обе стороны *подписали (Past)* меморандум о взаимопонимании в связи с этим также включает в себя сотрудничество в области создания центров для испытания новых инноваций в производство поездов».

The decision in this context would be the integration into the system of models based on the synthesis of the morphological models and the semantics of the used tokens.

3. Another problem of computer-aided translation is related to the principles of cross-language correspondence. We are talking about the fact that the information of the original text is firstly expressed in the neutral intermediate language, and then translated into the target language. The situation of this kind occurs in translation from Russian to Arabic and vice versa. It is known that typologically the Arabic and Russian languages are synthetic, while the English is analytical. Therefore, the use of English as the intermediary language brings errors to the grammatical meanings of words. Thus, the process of algorithms creating for correspondence of the morphological patterns in Russian and Arabic would help to enhance the quality of translation.

ولعل من أبرز نتائج التعاون، ازدياد عدد *الطلبة السعوديين* في الولايات المتحدة الأمريكية من حوالي ثلاثة آلاف طالب في السنوات القليلة الماضية، إلى نحو خمسة وعشرين الف طالب حاليا

«Perhaps the most important results of cooperation, the increasing number of *Saudi students* in the United States of about three thousand students in the past few years, to about twenty-five thousand students at present».

«Пожалуй, наиболее важных результатов сотрудничества, увеличение числа *Саудовской студентов* в Соединенных Штатах около 3000 студентов в последние несколько лет около 25 000 студентов в настоящее время».

It is necessary to note that sometimes we find mistakes in translation by means of so called "neutral language" (i.e. English in this case):

لن نخرج من مصر

«*Will not come out* of Egypt».

«*Не вышел* из Египта».

Thus, to ensure high-quality linguistic software perfection the structural formalization of the language must be flexible and comprehensive. The optimal development of linguistic software products requires application of comprehensive solution that includes the dictionary entries forming on the basis of frequency index, creation of the algorithms inflectional patterns of Arabic and also taking into account the semantic aspects.

«Perhaps the most important results of cooperation, the increasing number of Saudi students in the United States of about three thousand students in the past few years, to about twenty-five thousand students at present».

5 Conclusions

5.1 Considering the complexity and multifaceted character of the problem of the Arabic text formalization and OC recognition, the efforts of the specialists of different specializations - linguists (selection of the linguistic content, problems of the linguistic data parsing) and applied mathematicians (image and voice recognition, filtering, adaptive systems and software design) should be joined.

5.2 The effectiveness of the search engine should be based on automatic methods of the lexical material formalization, i.e. the definition the initial word and the root morpheme from the variety of written forms. Since in Arabic the root morphemes play the most important role as the 'holder' of the basic meanings and ideas, the lemmatization is also important for building of electronic dictionaries and automatic translation.

5.3 There is a big number of written variants of Arabic verbs and nouns which depends on grammatical categories of definite / indefinite, case (nominative, genitive, accusative), number and written realization (absence of the realization) of some graphic elements: such as dots under "yā' - ي ", hamza,

madda, waşla, tashdeed, etc., i.e. the elements which have facultative character in modern Arabic texts, especially in printed ones. So the number of the written variants is very big. The lowest number have separable prepositions, the highest variety of the written forms have inseparable particles and prepositions. In other words, those particles may be written with the majority of the noun or verbal forms.

Thus, the initial task is to develop automatic methods of the formalization of the lexical material and derive from the variety of word forms the vocabulary word form, and, finally, the root morpheme.

It will enable to create the linguistic content of the electronic dictionary on the basis of written and electronic Arabic texts, and the main criteria here is the frequency of the word entering into the text. It allows us to include the most frequent words into the linguistic content of the dictionary, and optimize the content of the dictionary.

The electronic dictionary based on the principle of the frequency code is very important, especially on the initial stages of studying Arabic.

As a result computerized methods of the definition of the frequency code of the Arabic words have been developed (the number the word entrances in the entire text on the basis of the analysis of about one million Arabic words included into vast variety of texts).

Table 1. Words in random with frequency code on the basis of analysis of Arabic written texts (about one million words).

شمس	1365
هذه	1365
إن	1275
عام	1274
ل	1271
مع	1175
ذلك	1096
علي	1089

Except for the limited number of dictionaries, the authors do not explain or stipulate the principles of the formation of the linguistic content and reflect authors' personal affiliations and criterion.

On the contrary, the words frequency code was the main criterion in the process of building the lexical content of the new Arabic-Russian dictionary.

Thus, the basic principles of the description of Arabic typical structures were developed along with the electronic Russian-Arabic and Arabic-Russian dictionaries.[8]

The principles of Arabic text processing were developed in cooperation with the team of mathematicians from St. Petersburg State University, among them O.Granichin and D Shalymov, who developed the implementations of various randomized algorithms of the stochastic multidimensional optimization such as SPSA (Simultaneous Perturbation Stochastic Approximation).[9] These principles may be used for the adaptive classification of the images and signals in the framework of uncertainty.

5.4 The research of the existing methods of the linguistic analysis, along with the positive experience in compilation of a set of morphological paradigms and dictionaries of Arabic, allows to develop a language model, which would be able to become the background for the development and improvement of machine translation technology, search engines, as well as the creation if linguistic software, and, on the other hand, it would contain the maximum possible amount of information for linguistic research. The proposed model focuses on the interaction between the morphological base, thematically tagged dictionary and text corpus. Each of these elements carries its functional load, however, only their complex integration could contribute to the improvement of technological solutions for Arabic.

5 References

[1] Bar-Hillel, Y., "The Present Status of Automatic Translations of Languages"; Advances in Computers, Vol.1, 91-163, 2006.

[2] Bernikova O.A. "The Arabic Language as a Tool for Informational Systems Providing (in Russian)"; Proceedings of the XXVI International Conference on Source Studies and Historiography of Asia and Africa "Modernization and tradition", St.Petersburg, Russia, 346-347, April 2011.

[3] Bernikova O.A. "The Arabic Grammar in Tables and Diagrams (in Russian)"; St.Petersburg, Russia, 2010.

[4] Bernikova O.A. "Problems of the Arabic Language Formalization (in Russian)"; Proceedings of the Conference "Oriental Studies and Dialog of Civilizations", St.Petersburg, Russia, 371-372, April 2009.

[5] Bernikova O.A. "Oriental Languages Distance Learning System (in Russian)"; Saint-Petersburg State University, Vol.17, 17-19, 2007.

[6] Granichin O.N., Vakhitov A.T., Gurevich L.S. "Algorithm for Stochastic Approximation with Trial Input

[8] Redkin O.I., Bernikova O.A., Shalymov D.S. "Software for Arabic text e-learning, translation and recognition"; St.Petersburg, Russia, April 2007. Certificate of the Official Software Registration No. 2007611711 dated 23.04.2007 «Software for Arabic Text e-Learning, Translation and Recognition», Russian Federation IP Agency.

[9] Vakhitov A.T., Granichin O.N., Sysoev S.S. Accuracy of Randomized Algorithm for Stochastic Approximation // Automation and Remote Control, 2006, v.67, No. 4, pp.589-597.

Perturbation in the Nonstationary Problem of Optimization"; Automation and Remote Control, Vol. 70, no. 11, 1827-1835, 2009.

[7] Granichin O.N., Hieu L.T. "Using Application of Statistics for Word Extraction from Vietnamese documents" ; Vestnik Sankt-Petersb. Univ. App. Math, no. 3, pp. 162-170, 2009.

[8] Granichin O.N., Vakhitov A.T., Gurevich L.S., "Automation and Remote Control", 2009.

[9] Granichin O.N., Polyak B.T. "Randomized Algorithms of an Estimation and Optimization Under Almost Arbitrary Noises". Nauka, Moscow, 2003.

[10] Izwaini S., "Problems of Arabic Machine Translation"; Proceedings of the International Conference on the Challenges of Arabic Machine Translation and Natural Language Processing, British Computer Society, London, 118-148, 23 October 2006.

[11] Mesfar S. "Standard Arabic Formalization and Linguistic Platform for its Analysis"; Proceedings of the Conference The Challenge of Arabic for NLP/MT, London, England, 84-95, 2006.

[12] McCarthy John J. "A Prosodic Theory of Nonconcatenative Morphology"; Linguistic Inquiry, Vol. 12, 373-418, 1981.

[13] Koskenniemi, Kimmo, "Two-level Morphology: A General Computational Model for Word-Form Recognition and Production, Publications", University of Helsinki, Department of General Linguistics, 1983.

[14] Redkin O.I. "Classical School of Oriental and Digital Technologies - New Opportunities (in Russian)"; Proceedings of the XXVI International Conference on Source Studies and Historiography of Asia and Africa "Modernization and tradition", St.Petersburg, Russia, 367-368, April 2011.

[15] Redkin O.I. "Formation of the Thesaurus of Arabic on the Basis of Frequency (in Russian)"; Proceedings of the Conference "Oriental Studies and Dialog of Civilizations", St.Petersburg, Russia, 379-380, April 2009.

[16] Redkin O.I. , Bernikova O.A., "Applying of technical equipment in teaching of the Arabic language"; Asian and African Studies, St.Petersburg, 124, 2006.

[17] Redkin O.I., Bernikova O.A., Shalymov D.S. "Software for Arabic text e-learning, translation and recognition"; St.Petersburg, Russia, April 2007.

[18] Shalymov D S. "Continuous Speech Recognition Using Simultaneous Perturbation Stochastic Approximation Algorithm"; Vestnik St.Petersburg State University, Vol.3, 171-181, 2009.

[19] Shalymov D S. "Automatic Recognition of the Printed Arabic Texts"; Stochastic Optimization in Computer Science, Vol.3, 124-137, 2009.